Credit Card Fraud Detection

Ali Rifat Kaya

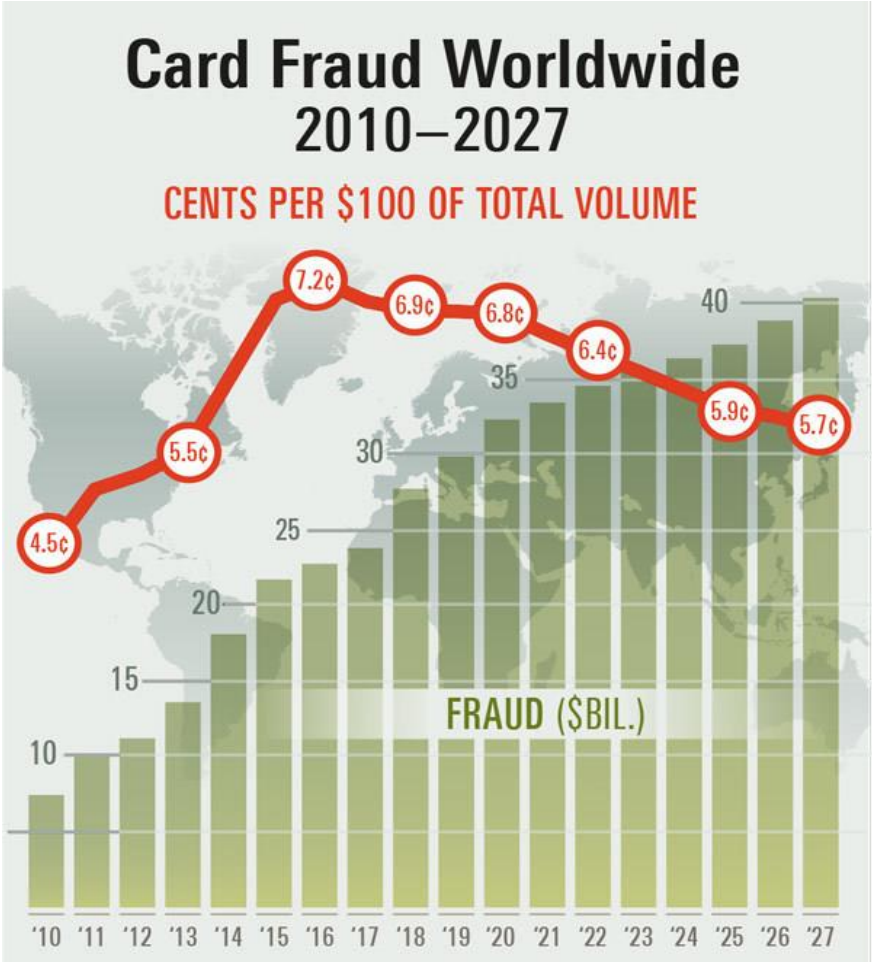
Fraudulent Transactions

For every 100$ genuine transaction, there is a 5.7 ¢ fraudulent transaction.

Does it seem insignificant ?

The amount of fraudulent transactions is expected to reach **40 B. $** in 2027.

Source: The Nilson Report, November (2019)

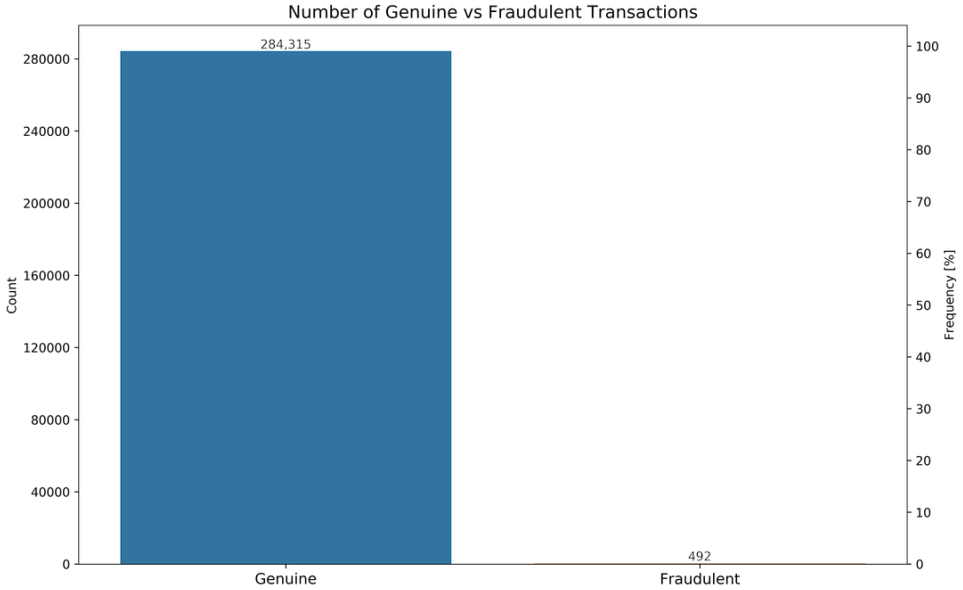


Fraud Detection

- Ideally, businesses want to find ways to prevent fraud from taking place, or, if that's not possible, **to detect it before significant damage is done**.
- Fraud detection occurs during the fraud attempt.
- The goal of fraud detection is to mitigate fraud.
- Sophisticated fraud detection solutions also reduce false positives which improves the user experience and increases the productivity of fraud teams
- However, the key point is to teach customers how to stay safe:
  - **- Use Safe Payments**
  - **- Guard Personal Information**
  - **- Check Credit Reports Regularly**
  - **- Conduct Online Searches**
  - **- Remain Skeptical About Free Trials**

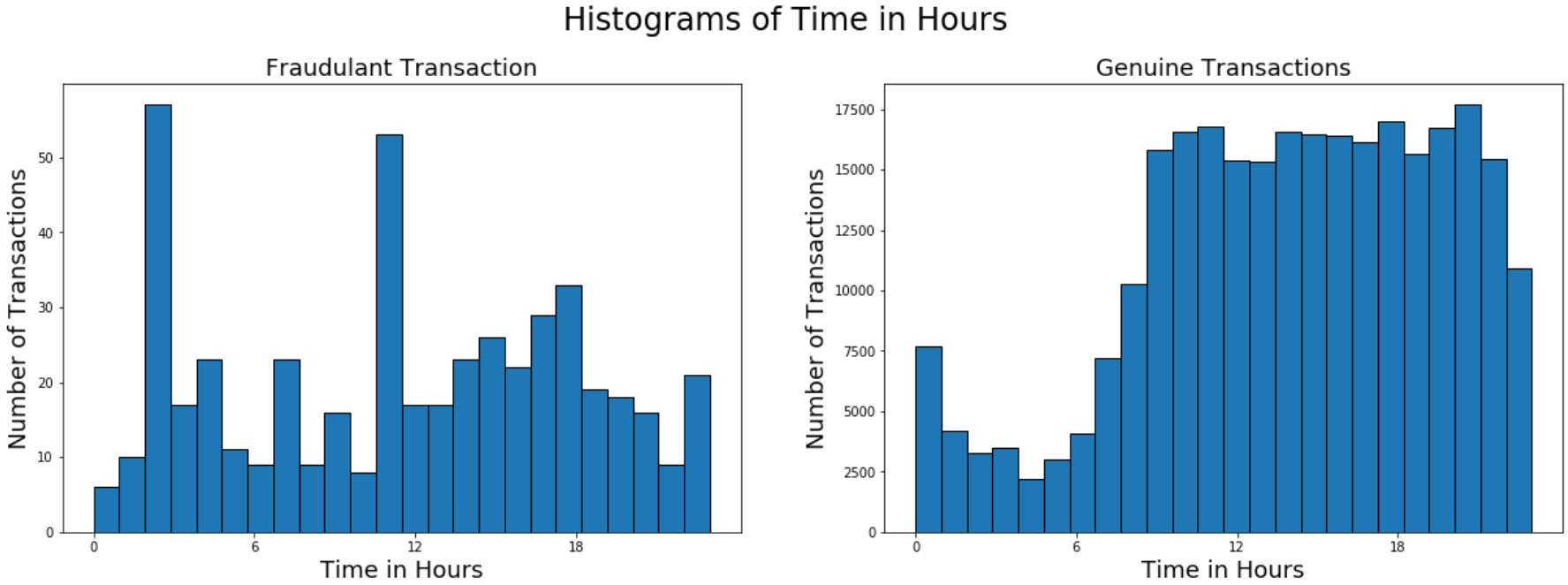
The Data

- 2 Days of transaction
- Collected from Europe
- Accounted for privacy
- 284.407 credit card transactions
- **Fraudulent to genuine transactions ratio 1:578**
- The genuine transactions account for **99.83%** of all transactions
- The fraudulent transactions only account for **0.17%** of all transactions



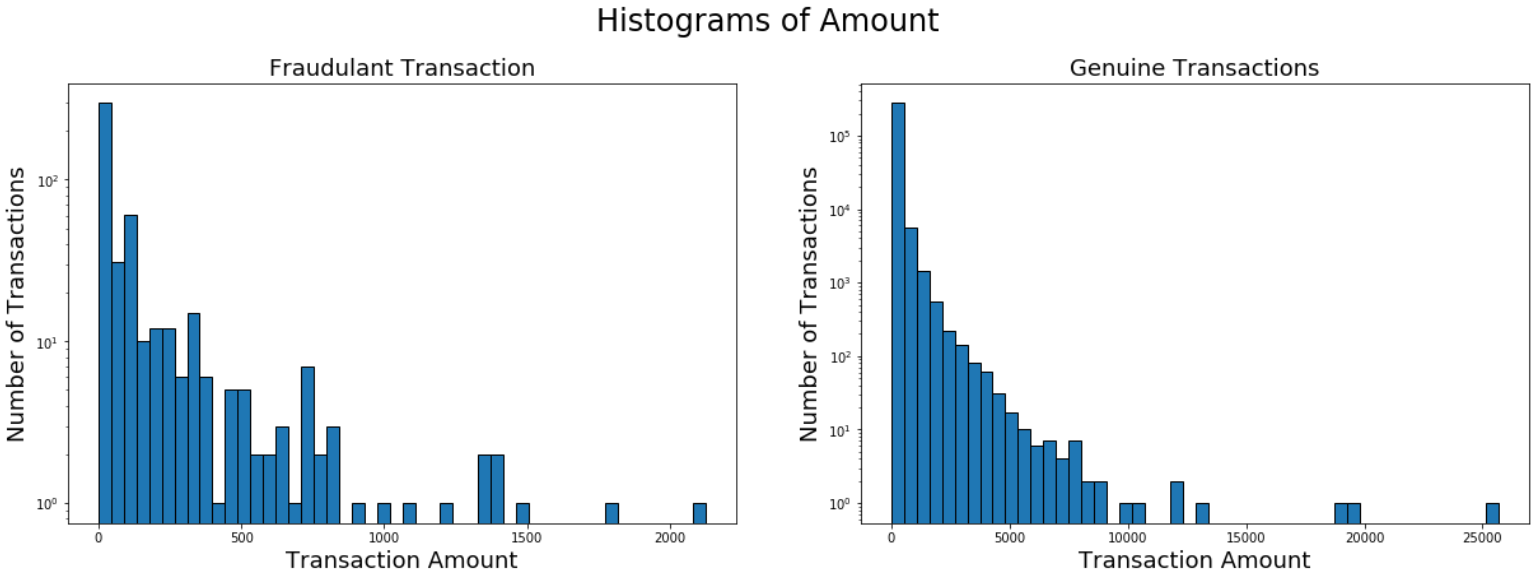
The Data (Continued)

- When transactions take place?



The Data (Continued)

- What are the amounts?



Fraudulent transactions mean amount: 122.211
Fraudulent transactions median amount: 9.250

Genuine transactions mean amount: 88.291
Genuine transactions median amount: 22.000

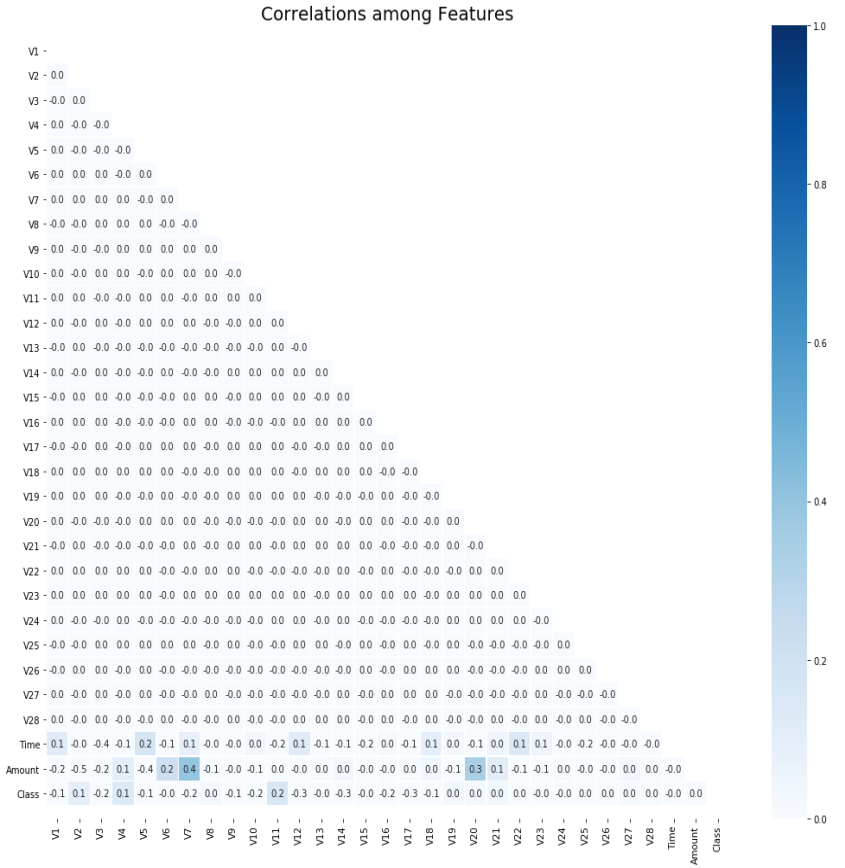
The Data (Continued)

- How to detect fraudulent transactions?

Correlation Analysis

There are no patterns among the features (input variables).

Thanks to PCA!!!



The Data (Continued)

- How to detect fraudulent transactions?

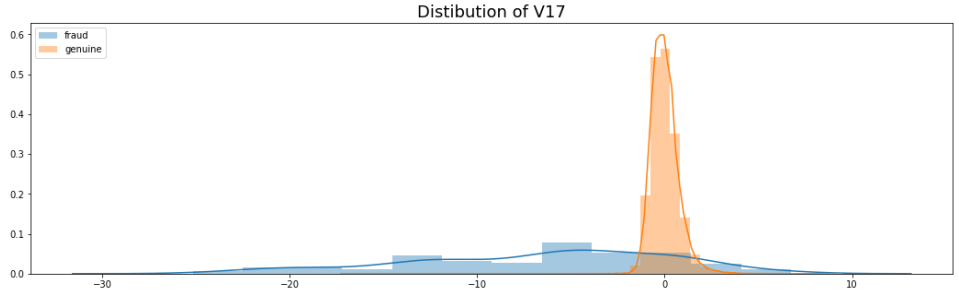
A Smarter Way

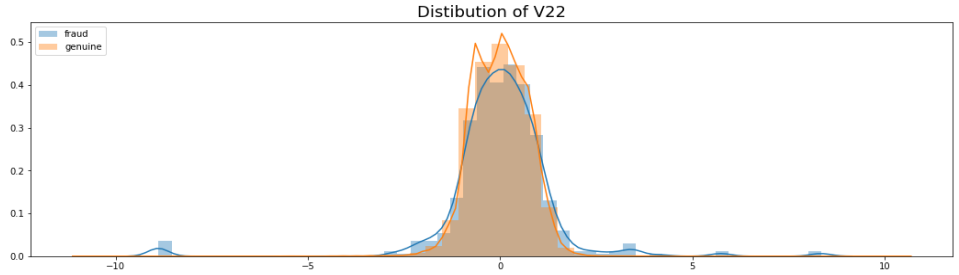
Bivariate Analysis

Explore each feature by transaction class:

- Genuine
- Fraudulent

Find the different





The Analysis

- Models implemented:
    - Logistic Regression
    - K-Nearest Neighbours
    - Naive Bayes
    - Decision Tree Classifier
    - Random Forest Classifier
    - Extra Trees Classifier
    - AdaBoost Classifier
    - Gradient Boosting Classifier
    - XGBoost Classifier

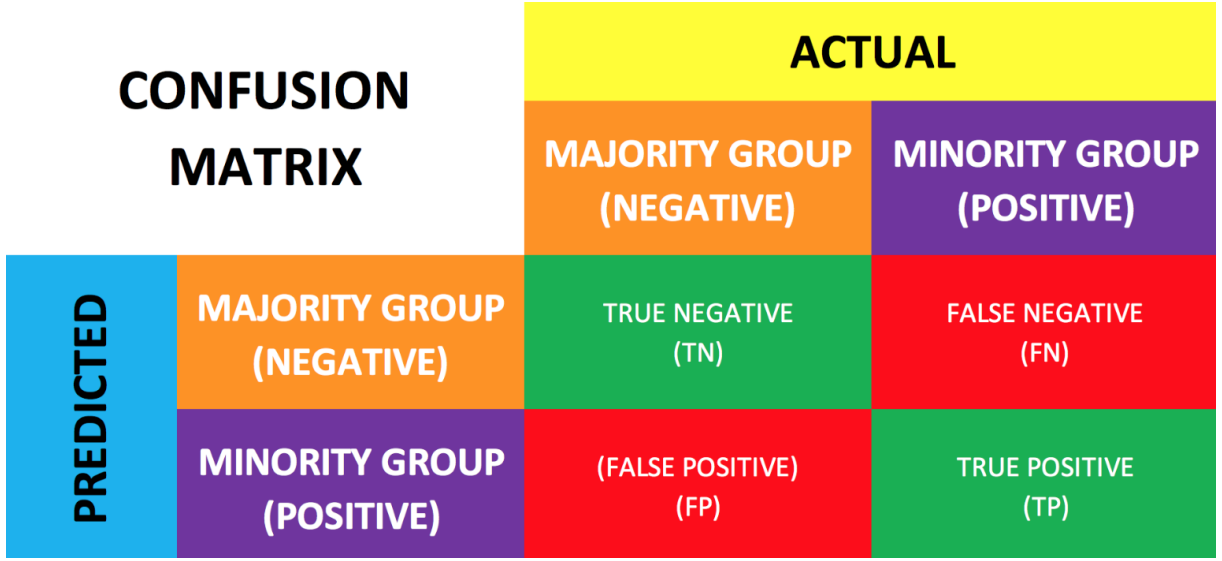
TOO MANY MODELS ?

Don’t worry - It’s just

The Analysis (Continued)

- The Confusion Matrix - the mother of all metrics

  Note: It can be really confusing!!!



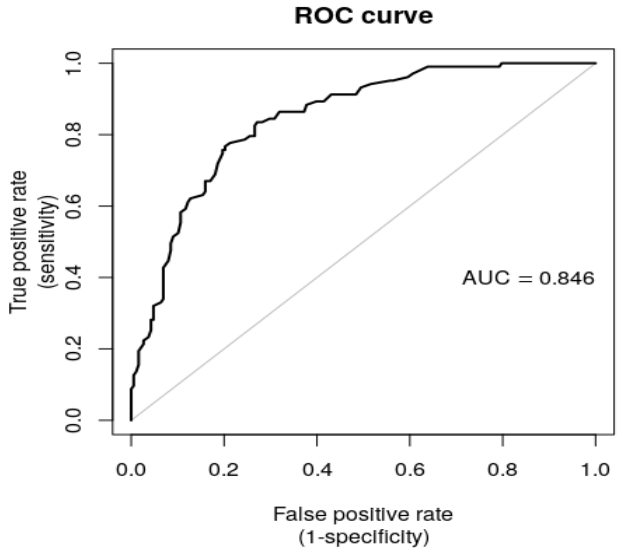
TN: Transactions that are **genuine** and predicted to be **genuine**

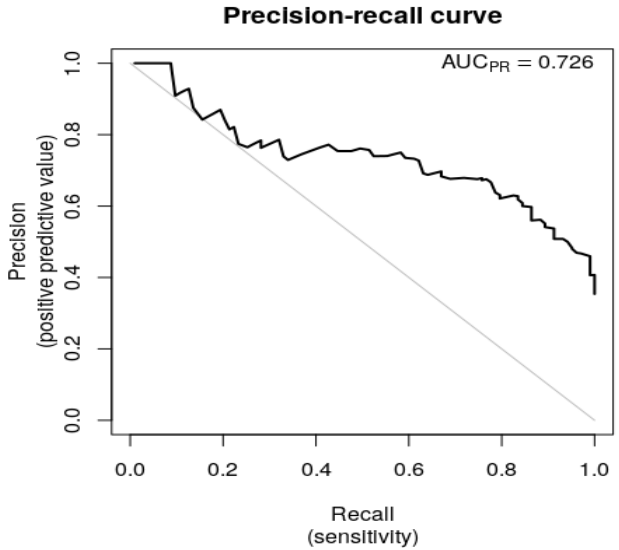
FN: Transactions that are **fraudulent** but predicted to be **genuine**

FP: Transactions that are **genuine** but predicted to be **fraudulent**

TP: Transactions that are **fraudulent** and predicted to be **fraudulent**

The Analysis (Continued)



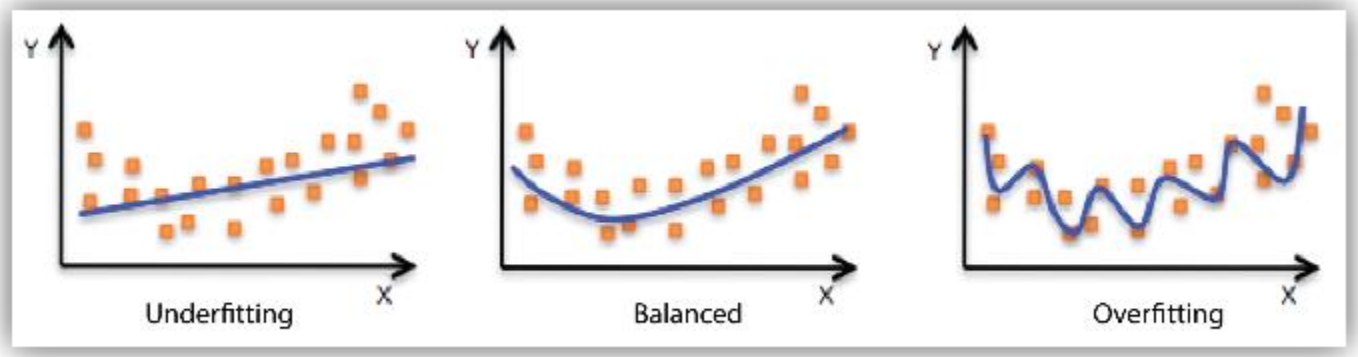


- ➔ Easy to see the trade-off:
  - ◆ Detecting all fraudulent transactions may not be the best answer

- ➔ Easy to interpret:
  - ◆ Higher score better model

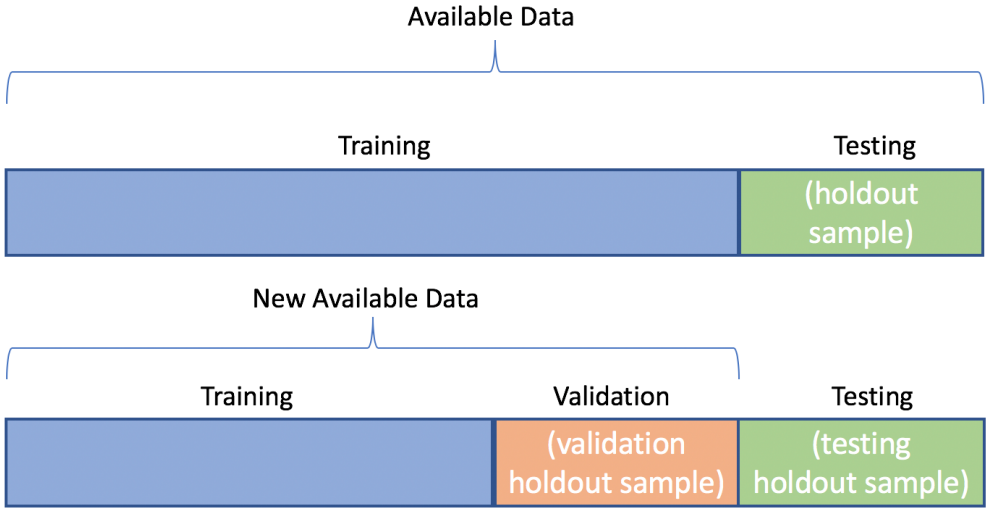
The Analysis (Continued)

- Overfitting and Underfitting - finding the balance
  - **Overfitting**: the model works only well with the data in hand
  - **Underfitting**: the model does not successfully capture the patterns in the data

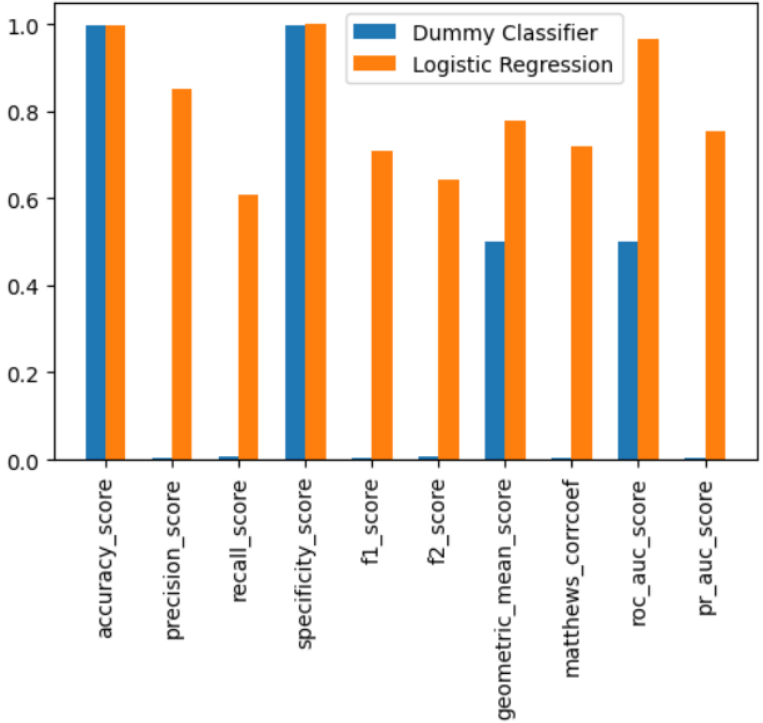


The Analysis (Continued)

- **Training Data**: The data that is used to train the model (Teach the patterns)
- **Test Data**: The data that the model is tested (How good the model learn the patterns?)
- **Cross-Validation**: split the training data into *k*-folds and use *k-1* folds (training data) to train the model and 1 fold (validation data) to validate the model.

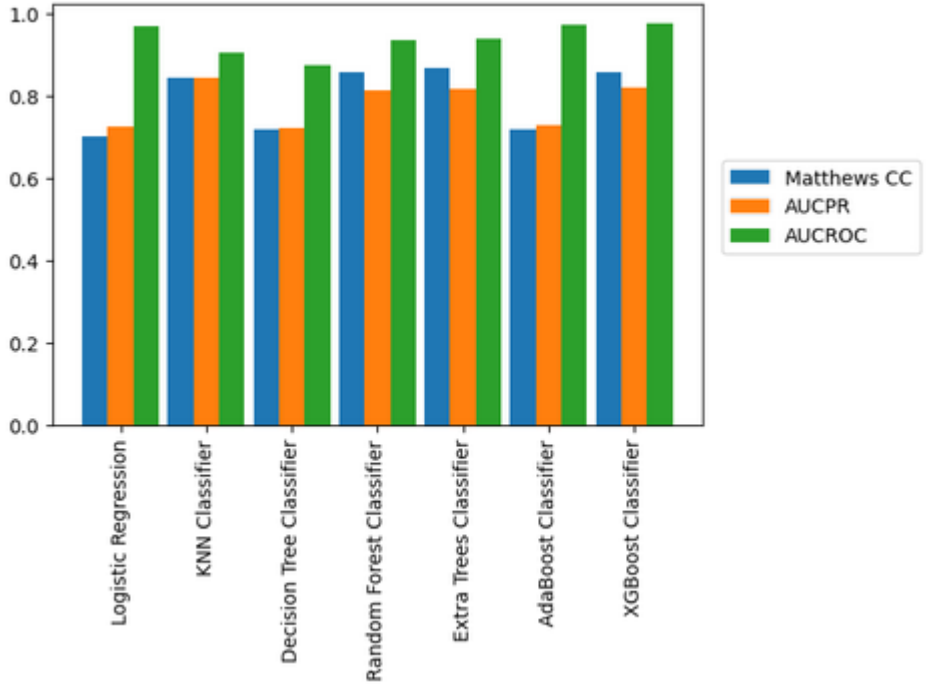


Machine Learning – Spot Check on Metrics



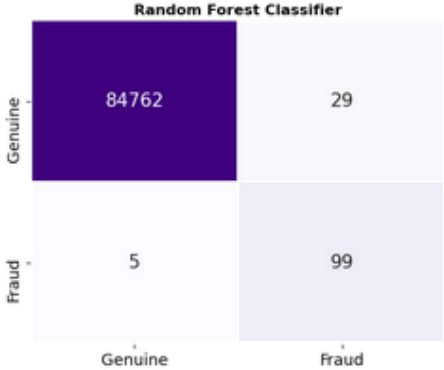
- Some well-known metrics may not work all the time
- It is good to start with a set of metrics but better is to stay with a small set of them
- MCC, AUCPR and AUCROC are the best for our case

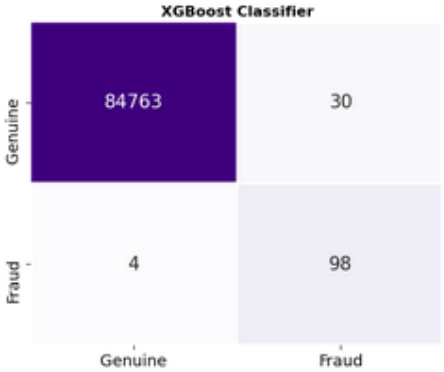
Machine Learning – First Results



- Summary of test results
- Shows how the models are doing
- The green bar (AUCROC) is inflated
- Select the best performing models:
    1. Extra Trees
    2. XGBoost
    3. Random Forest

Machine Learning - What about transactions?





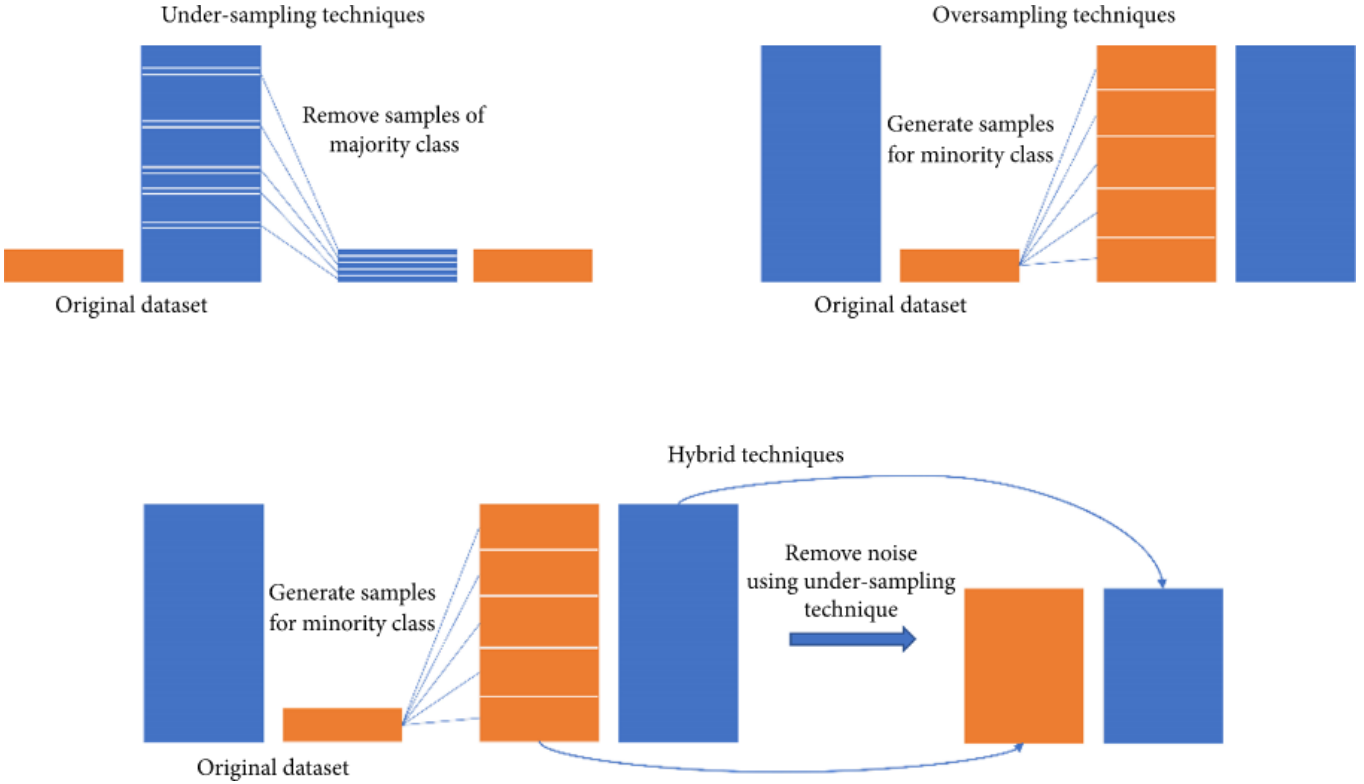
- High performance on identifying the genuine transactions.
- The false negatives (wrong alarms) are a manageable number of transactions.
- Look to improve false positives

Machine Learning - Sampling

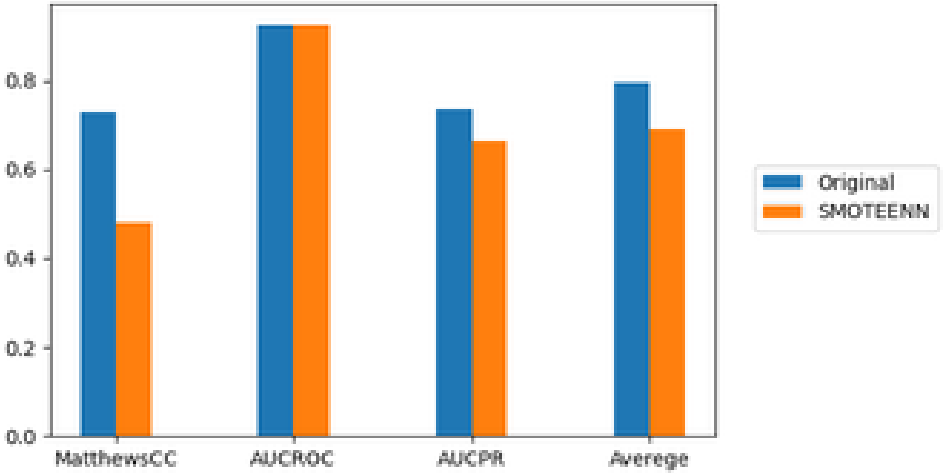
- **Sampling**: Changing the class distribution in order to improve the performance of the machine learning models
    - **Under-Sampling**: Deleting observations from the majority class (genuine transactions) in order to close the gap in the class distributions (the number of observation in each class)
    - **Over-Sampling**: Increasing the number of observations in the minority class in order to close the gap in the class distributions
    - **Hybrid Sampling:** The combination of over and under-sampling.

In this study, Hybrid Sampling –SMOTEENN- is applied.

Machine Learning - Sampling (Continued)



Machine Learning - Sampling (Continued)



- Hybrid Sampling didn’t improve the performance
- It is best to go with original data

Machine Learning - Optimiztion

	Default Parameters	Tuned Parameters	Change
Logistic Regression	0.756	0.758	0.002
Decision Tree Classifier	0.734	0.761	0.027
Random Forest Classifier	0.854	0.860	0.060
Extra Trees Classifier	0.861	0.864	0.003
AdaBoost Classifier	0.761	0.829	0.068
Gradient Boosting Classifier	0.604	0.828	0.224
XGBoost Classifier	0.850	0.861	0.011
K-Nearest Neighbors Classifier	0.852	0.859	0.007

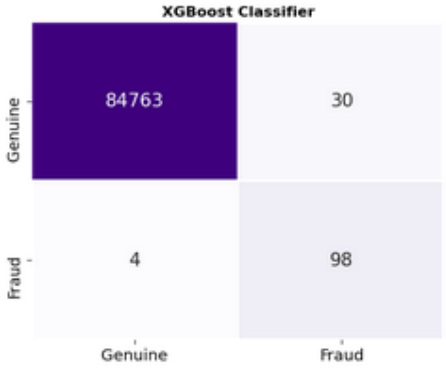
- Each model increased its performance.
- The greatest increase is 22 points
- The least increase is 0.2 points

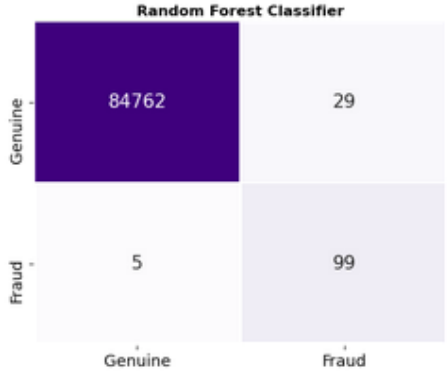
The best models are still didn't change.

- Extra Trees
- XGBoost
- Random Forest

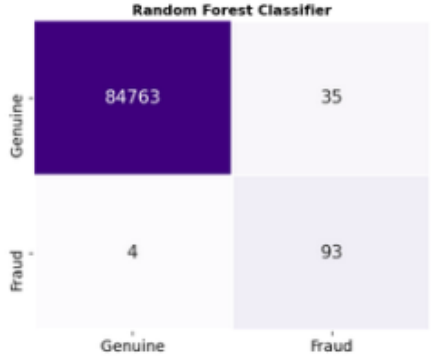
Final Evaluation

Default





Optimized



Conclusion

- Hyperparameter optimization, overall, did have a positive effect on the performance of algorithms. Based on the test scores and the confusion matrices the top 3 algorithms are:

1. XGBoost Classifier (Tuned Parameters)
2. Extra Trees Classifier (Default Parameters)
3. Random Forest(Default Parameters)

- Finally, the best algorithm is the XGBoost Classifier. It is able to catch 78.9% of the fraudulent transactions with less than 5 false negatives (0.00003%).